Parsing for 100 Arabic Sentences using Context Free Grammar

Kathrein Abu Kwaik

Formal Linguistics Course
CLASP
Göteborg University

22 november 2017

In this project we will employ Context Free Grammar (CFG) to parse small corpus for Modern standard Arabic that contains 100 sentences.

1 Challenging of Arabic Parsing

Natural Language processing (NLP) serves as the basic block for many Natural language applications like machine translation, speech processing and semantic analysis. NLP contains three fundamental component: lexicon, syntax and morphology. In this project we will work on Modern Standard Arabic (MSA) to show the hierarchical structure of this language. A small manually collected corpus contains 100 Arabic sentences are used to build the lexicon firstly before parsing. Arabic words classify into 3 categories : noun, verb and particle and the Arabic sentence can be only nominal sentence or verbal sentence. Arabic Syntax is a very complex field of study, even an Arabic native speakers are not fully familiar with the grammar of their native language. Parsing Arabic language consider a challenging task due to the following difficulties:

- Arabic is highly derivational and inflectional Language, where each component requires extensive study and exploitation of the associated linguistic characteristics. For example: Figure 1 lists 82 inflections for the same verb لعب *lEb* (play) based on their grammatical rule.

- Sentence Structure: MSA is syntactically ambiguous because of the frequent usage of grammatical relations, order of words and phrases and conjunctions.

الضمائر	المضارع المنصوب	الأمر المؤكد	الماضي المعلوم	المضارع المؤكد الثقيل	المضارع المعلوم	المضارع المجزوم	الأمر
أنا	أَلْعَبَ		لَعَبْتُ	أَلْعَبَنَّ	أَلْعَبُ	أَلْعَبْ	
نحن	نَلْعَبَ		لَعَبْنَا	نَلْعَبَنَّ	نَلْعَبُ	نَلْعَبْ	
أنت	تَلْعَبَ	اِلْعَبَنَّ	لَعَبْتَ	تَلْعَبَنَّ	تَلْعَبُ	تَلْعَبْ	اِلْعَبْ
أنتِ	تَلْعَبِي	اِلْعَبَنَّ	لَعَبْتِ	تَلْعَبَنَّ	تَلْعَبِينَ	تَلْعَبِي	اِلْعَبِي
أنتما	تَلْعَبَا	اِلْعَبَانِّ	لَعَبْتُمَا	تَلْعَبَانِّ	تَلْعَبَانِ	تَلْعَبَا	اِلْعَبَا
أنتما مؤ	تَلْعَبَا	اِلْعَبَانِّ	لَعَبْتُمَا	تَلْعَبَانِّ	تَلْعَبَانِ	تَلْعَبَا	اِلْعَبَا
أنتم	تَلْعَبُوا	اِلْعَبُنَّ	لَعَبْتُم	تَلْعَبُنَّ	تَلْعَبُونَ	تَلْعَبُوا	اِلْعَبُوا
أنتن	تَلْعَبْنَ	اِلْعَبْنَانِّ	لَعَبْتُنَّ	تَلْعَبْنَانِّ	تَلْعَبْنَ	تَلْعَبْنَ	اِلْعَبْنَ
هو	يَلْعَبَ		لَعَبَ	يَلْعَبَنَّ	يَلْعَبُ	يَلْعَبْ	
هي	تَلْعَبَ		لَعَبَتْ	تَلْعَبَنَّ	تَلْعَبُ	تَلْعَبْ	
هما	يَلْعَبَا		لَعَبَا	يَلْعَبَانِّ	يَلْعَبَانِ	يَلْعَبَا	
هما مؤ	تَلْعَبَا		لَعَبَتَا	تَلْعَبَانِّ	تَلْعَبَانِ	تَلْعَبَا	
هم	يَلْعَبُوا		لَعَبُوا	يَلْعَبُنَّ	يَلْعَبُونَ	يَلْعَبُوا	
هن	يَلْعَبْنَ		لَعَبْنَ	يَلْعَبْنَانِّ	يَلْعَبْنَ	يَلْعَبْنَ	

Figur 1: 82 inflections for an verb play

- The lack of resource like large manually tagged Corpus.
- The Arabic sentences normally are too long in terms of number of words.
- The omission of Diacritics (vowel) in written Arabic which can change the meaning of the words and it's POS tag.
- The free word order in Arabic Sentence, the most common order is VSO, and there are also SVO, OVS and VO.
- The presence of the Elliptic personal pronoun as (كتب الدرس *ktb āldrs*: He wrote the lesson)
- The POS tag in some cases is based on the word it self and relates to some words surround it, like the absolute object where it is used to emphasize the verb: for example in this two sentences:

  1. ضحكت ضحكة *ḍḥkt ḍḥkh*: I laughed a laugh
  2. ضحكت مبتسمة *ḍḥkt mbtsmh*: I laughed with a smile

  Even thought the two words ضحكة *ḍḥkh* and مبتسمة *mbtsmh* share the same position and meaning, but the tag is different for each of them. In the first sentence it is absolute object while it is an adverb in the second sentence.

- Arabic need huge morphological segmentation works before applying parser because in most cases the Arabic letters and affixes joint together to form single words, for example (وسيتكتبونها *wsytktbwnhā* and they will write it).

2 Corpus description

This is a 100 Arabic sentence corpus which is written manually and it covers the main phenomena in Arabic Grammar. The corpus contains 327 word with 225 tokens. Given that Arabic sentence can be Verbal or nominal, so we starts to cover the main types of verbal sentences (57 sentences) then we added some cases of nominal grammar(43 sentences).

2.1 Verbal Sentences (VP)

The most common structure for the VP is VSO, and the main basic form of VP is the V with personal pronoun like (كتب *ktb* write). The verb indicates the person, number and gender of the subject and in some cases the object. we will list some forms of VP:

- Verb with Pronoun as subject: (كتبنا *ktbnā* we write)
- Verb with nominative Noun as subject (كتب الولد الأولاد *ktb ālwld āl'awlā-d* The boy/boys wrote)
- Verb with elliptic Subject and attached Object (كتبها *ktbhā* he wrote it)
- Verb with attached Object and nominative Subject (كتبهاالولد *ktbhāālwld* the boy wrote it)
- Verb with elliptic Subject and accusative Object (كتب القصة *ktb ālqṣh* he wrote the story)
- Verb with nominative Subject and accusative Object (كتب الولد القصة *ktb ālwld ālqṣh* the boy wrote the story)
- Verb with nominative Subject, first accusative Object and second accusative Object (أعطى الولد البنت كتابا *'aṭā ālwld ālbnt ktābā* the boy gave the girl a book)

- Verb with nominative Subject, second accusative Object + to + first genitive Object:(أعطى الولد كتابا للبنت *ʾaṭā ālwld ktābā llbnt* the boy gave a book to the girl)
- verb with attached subject, attached first object and second accusative Object (أعطيناها كتابا *ʾaṭynāhā ktābā* we gave her a book)
- verb with attached first object, nominative Subject and second accusative Object. (أعطاها الولد كتابا *ʾaṭāhā ālwld ktābā* he gave her a book)
- (Kana and her sisters) verb with nominative noun and accusative noun (كانت الشمس ساطعة *kānt ālšms sāṭʿh* The sun was bright)
- (Kana and her sisters) verb with nominative noun and nominal sentence (كان الفلاح عمله شريف *kān ālflāḥ ʿmlh šryf* The farmer was a decent work)
- (Kana and her sisters) verb with nominative noun and verbal sentence (كان المعلم يدرس التلاميذ *kān ālmʿlm ydrs āltlāmyḏ* The teacher taught the students)
- (Kana and her sisters) verb with prepositional phrase and nominative noun (ليس للخائن وطن *lys llẖāʾyn wṭn* No homeland for the traitor)

2.2 Nominal Sentences (NP)

The Nominal Sentence (NP) has the form of Topic and Complement. The following lists some of the NP form:

- Definite nominative noun with Indefinite nominative noun (الشمس ساطعة *ālšms sāṭʿh* The sun is bright)
- N-Sent: Definite nominative noun with preposition phrase (الرجل في البيت *ālrğl fy ālbyt* The man is in the house)
- N-Sent: Definite nominative noun with nominal sentence (البيت بابه جديد *ālbyt bābh ğdyd* The house, its door is new)
- Definite nominative noun with verbal sentence (الأولاد كتبوا القصص *āl-ʾawlād ktbwā ālqṣṣ* The boys wrote the stories)
- (Ena and her sisters) with accusative noun and nominative noun (إن المعلمين شريفون *ʾin ālmʿlmyn šryfwn* The two teachers are honorable)

In addition to the previous examples we cover also the adjective cases, the adverb, the subjective pronouns, the Demonstrative pronouns, (single, dual and plural grammar) and the relative clauses phenomena.

3 Parsing with CFG

Although Arabic parsing applications should apply morphological segmentation before, but we didn't do that. we apply parsing directly to the text as we don't need to build complicated system. Context free grammar with features agreement are implemented to parse the 100 sentences corpus. Most sentences have only one parse tree as it describes one rule in Arabic grammar, while the other may have more than 3 parse tree as a complicated sentence describes more than one rule.